Introduction to XML and Digital Scholarly Editing Using the Text Encoding Initiative (TEI)

Paul Broyles, North Carolina State University
pabroyle@ncsu.edu

Workshop Outline

I. Overview of Digital Editions
II. XML and Markup
III. The TEI
IV. Hands-On Text Encoding
V. Where do we go from here?

Overview of Digital Editions

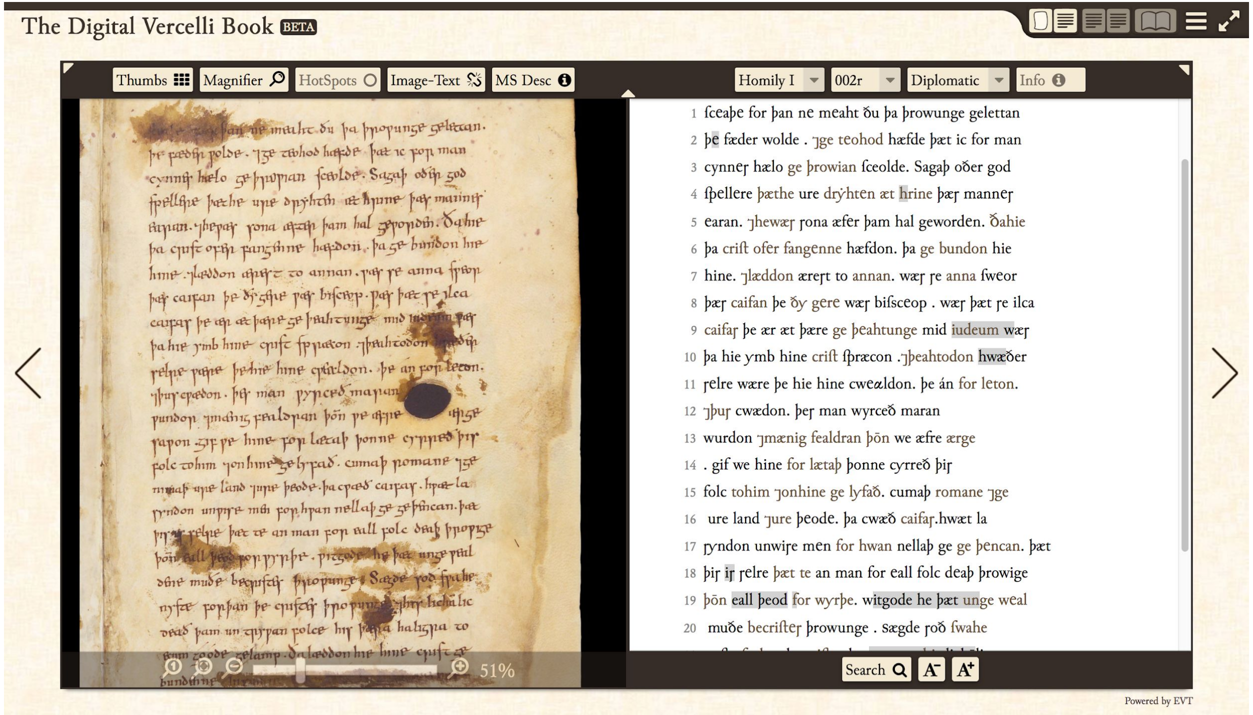
What is an edition?

- Edition: a particular form in which a work is published; the set of copies of a book produced from a single setting of type.
- Scholarly edition: remediates or represents documents or texts in a form determined by scholarly inquiry.
    - Critical edition: reconstructs a hypothetical text (e.g. a lost copy, an author's original, or an ideal text that never existed) using evidence such as multiple copies, linguistics, etc.
    - Documentary edition: "an edition of a text based on a single document, which attempts to reproduce a certain degree of the peculiarities of the document itself" (Pierazzo, "Digital Documentary Editions and the Others")
- Transcription: a reproduction of the characters that make up a text. *Transcription is interpretation; not neutral.*

Why digital editions?

- Online publication makes editions easily accessible; wide audience.
- Low publication costs make it economical to edit and publish material that might not be publishable in print; explosion of digital documentary editions.
- High degrees of flexibility and interactivity. No longer necessary to choose between documentary and critical editing; edition can contain both. Innovative forms of representation.
- Ability to tie edited texts to facsimile images of original source material.
- Collaboration and ongoing improvements.
- Makes text available as (structured) data for analysis, visualization.
- Multiple output formats: create encoded digital text once, use in a variety of systems for different purposes.

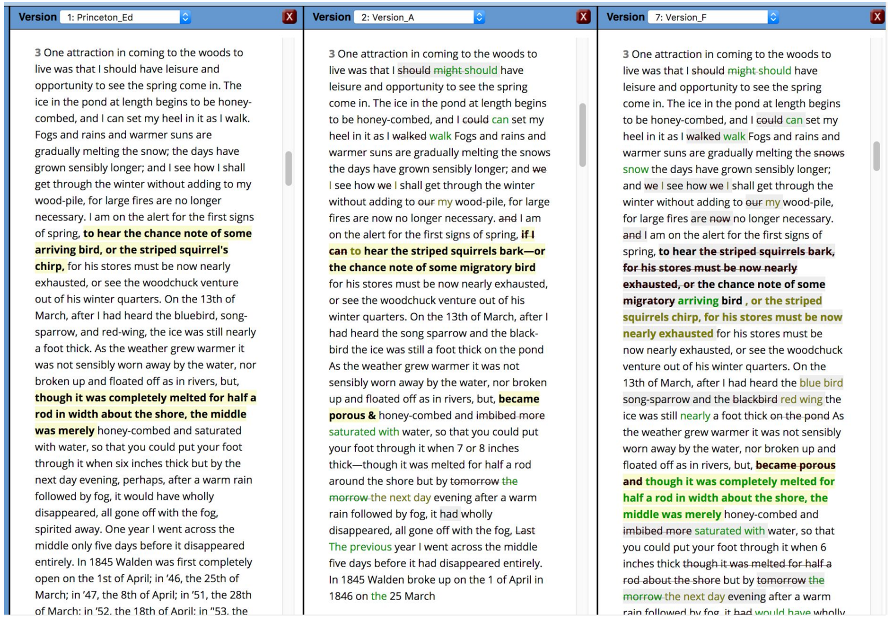
Represent primary sources



ſceaþe for þan ne meaht ðu þa þrowunge gelettan
þe fæder wolde . ⁊ge teohod hæfde þæt ic for man
cynneꞅ hælo ge þrowian ſceolde. Sagaþ oðer god
ſþellere þæthe ure drẏhten æt hrine þæꞅ manneꞅ
earan. ⁊hewæꞅ ꞅona æfer þam hal geworden. ðahie
þa criſt ofer fangenne hæfdon. þa ge bundon hie
hine. ⁊læddon æreꞅt to annan. wæꞅ ꞅe anna ſweor
þæꞅ caifan þe ðy gere wæꞅ biſsceop . wæꞅ þæt ꞅe ilca
caifaꞅ þe ær æt þære ge þeahtunge mid iudeum wæꞅ
þa hie ymb hine criſt ſþræcon .⁊þeahtodon hwæðer
ꞅelre wære þe hie hine cweαldon. þe án for leton.
⁊þuꞅ cwædon. þeꞅ man wyrceð maran
wurdon ⁊mænig fealdran þōn we æfre ærge
. gif we hine for lætaþ þonne cyrreð þiꞅ
folc tohim ⁊onhine ge lyfað. cumaþ romane ⁊ge
ure land ⁊ure þeode. þa cwæð caifaꞅ.hwæt la
ꞅyndon unwiꞅe men for hwan nellaþ ge ge þencan. þæt
þiꞅ iꞅ ꞅelre þæt te an man for eall folc deaþ þrowige
þōn eall þeod for wyrþe. witgode he þæt unge weal
muðe becriſteꞅ þrowunge . sægde ꞅoð ſwahe

Digital Vercelli Book <http://vbd.humnet.unipi.it/beta2/>, a parallel facsimile and diplomatic edition of an Anglo-Saxon manuscript

Flexible critical texts

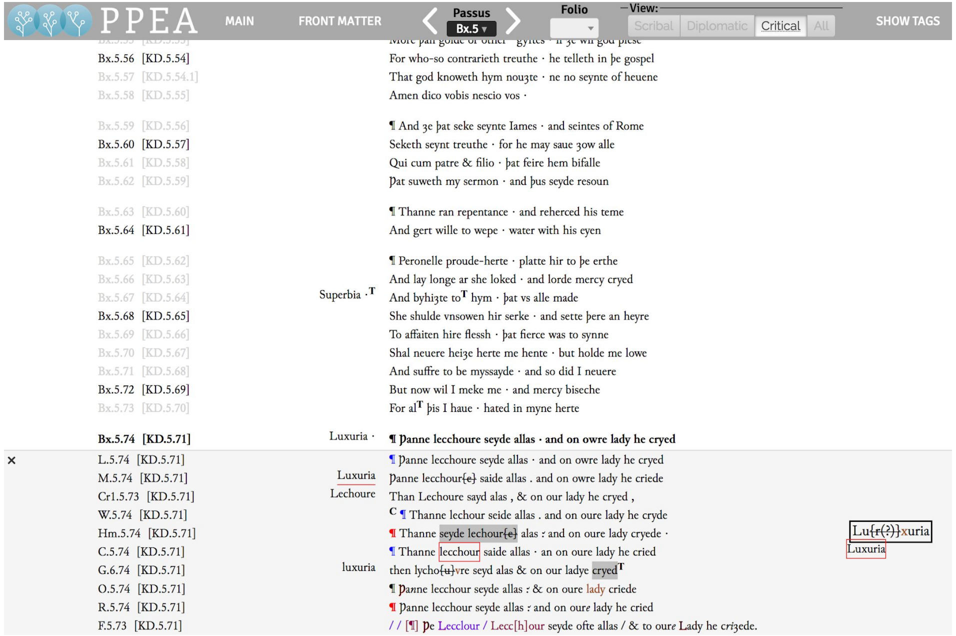


3 One attraction in coming to the woods to live was that I should have leisure and opportunity to see the spring come in. The ice in the pond at length begins to be honey-combed, and I can set my heel in it as I walk. Fogs and rains and warmer suns are gradually melting the snow; the days have grown sensibly longer; and I see how I shall get through the winter without adding to my wood-pile, for large fires are no longer necessary. I am on the alert for the first signs of spring, **to hear the chance note of some arriving bird, or the striped squirrel's chirp,** for his stores must be now nearly exhausted, or see the woodchuck venture out of his winter quarters. On the 13th of March, after I had heard the bluebird, song-sparrow, and red-wing, the ice was still nearly a foot thick. As the weather grew warmer it was not sensibly worn away by the water, nor broken up and floated off as in rivers, but, **though it was completely melted for half a rod in width about the shore, the middle was merely** honey-combed and saturated with water, so that you could put your foot through it when six inches thick but by the next day evening, perhaps, after a warm rain followed by fog, it would have wholly disappeared, all gone off with the fog, spirited away. One year I went across the middle only five days before it disappeared entirely. In 1845 Walden was first completely open on the 1st of April; in '46, the 25th of March; in '47, the 8th of April; in '51, the 28th of March; in '52, the 18th of April; in "53, the

3 One attraction in coming to the woods to live was that I ~~should~~ ~~might~~ should have leisure and opportunity to see the spring come in. The ice in the pond at length begins to be honey-combed, and I ~~could~~ can set my heel in it as I ~~walked~~ walk Fogs and rains and warmer suns are gradually melting the snows the days have grown sensibly longer; and ~~we~~ I see how ~~we~~ I shall get through the winter without adding to ~~our~~ my wood-pile, for large fires are now no longer necessary. ~~and~~ I am on the alert for the first signs of spring, **~~if I~~ ~~can~~ to hear the striped squirrels bark—or the chance note of some migratory bird** for his stores must be now nearly exhausted, or see the woodchuck venture out of his winter quarters. On the 13th of March, after I had heard the song sparrow and the black-bird the ice was still a foot thick on the pond As the weather grew warmer it was not sensibly worn away by the water, nor broken up and floated off as in rivers, but, **became porous &** honey-combed and ~~imbibed more~~ saturated with water, so that you could put your foot through it when 7 or 8 inches thick—though it was melted for half a rod around the shore but by ~~tomorrow~~ the ~~morrow~~ the next day evening after a warm rain followed by fog, it had wholly disappeared, all gone off with the fog, ~~Last~~ The previous year I went across the middle five days before it had disappeared entirely. In 1845 Walden broke up on the 1 of April in 1846 on the 25 March

3 One attraction in coming to the woods to live was that I ~~should~~ ~~might~~ should have leisure and opportunity to see the spring come in. The ice in the pond at length begins to be honey-combed, and I ~~could~~ can set my heel in it as I ~~walked~~ walk Fogs and rains and warmer suns are gradually melting the ~~snows~~ snow the days have grown sensibly longer; and ~~we~~ I see how ~~we~~ I shall get through the winter without adding to ~~our~~ my wood-pile, for large fires are ~~now~~ no longer necessary. ~~and~~ I am on the alert for the first signs of spring, **~~to hear the striped squirrels bark, for his stores must be now nearly exhausted, or~~ the chance note of some ~~migratory~~ arriving bird , or the striped squirrels chirp, for his stores must be now nearly exhausted** for his stores must be now nearly exhausted, or see the woodchuck venture out of his winter quarters. On the 13th of March, after I had heard the blue bird song-sparrow and the ~~blackbird~~ red wing the ice was still nearly a foot thick ~~on the pond~~ As the weather grew warmer it was not sensibly worn away by the water, nor broken up and floated off as in rivers, but, **~~became porous and~~ though it was completely melted for half a rod in width about the shore, the middle was merely** honey-combed and ~~imbibed more~~ saturated with water, so that you could put your foot through it when 6 inches thick ~~though it was melted for half a rod about the shore~~ but by ~~tomorrow~~ the ~~morrow~~ the next day evening after a warm rain followed by fog, it ~~had~~ would have wholly

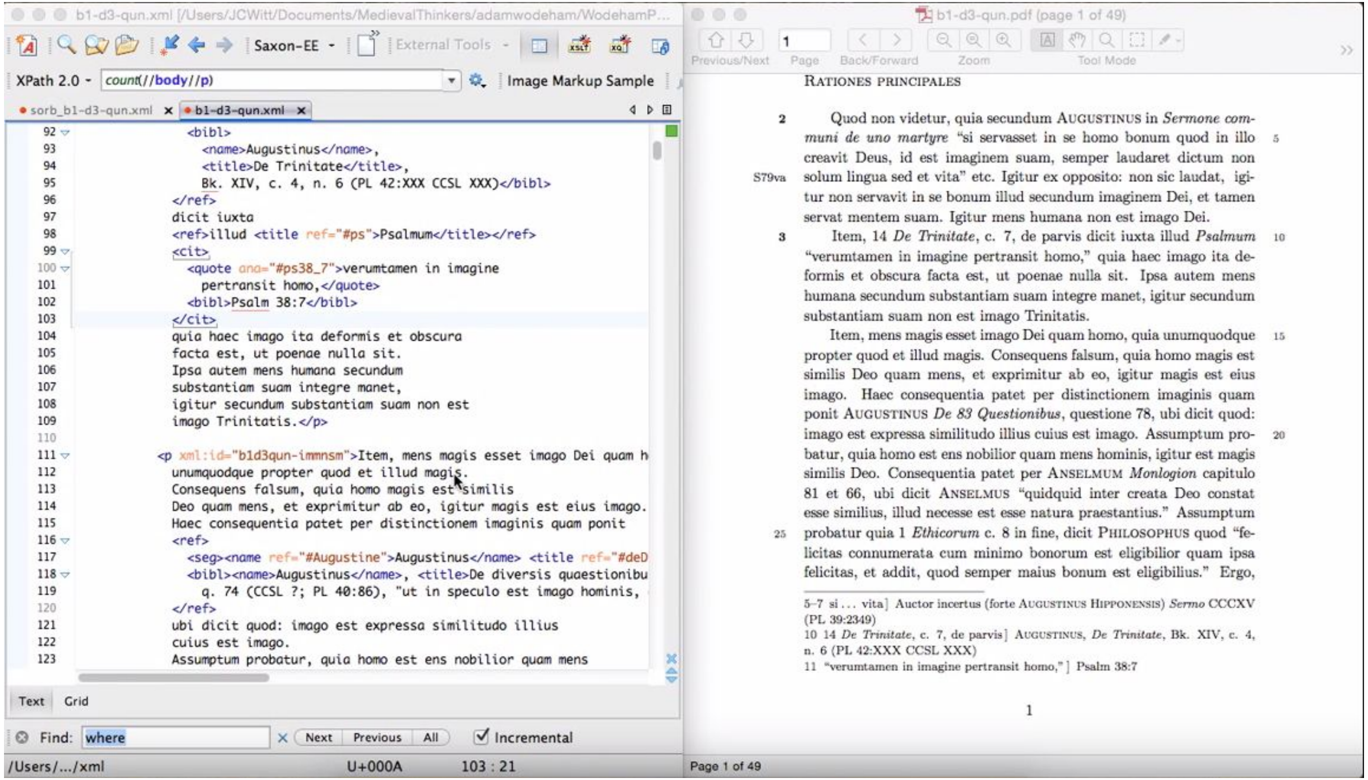
Comparing two manuscript versions of Thoreau's *Walden* with a standard print edition (left). *Walden: A Fluid-Text Edition*. Digital Thoreau. <https://digitalthoreau.org/fluid-text-toc/>.



Bx.5.55 [KD.5.53]		More þan golde or other · gyftes · if ȝe wil god plese	
Bx.5.56 [KD.5.54]		For who-so contrarieth treuthe · he telleth in þe gospel	
Bx.5.57 [KD.5.54.1]		That god knoweth hym nouȝte · ne no seynte of heuene	
Bx.5.58 [KD.5.55]		Amen dico vobis nescio vos ·	
Bx.5.59 [KD.5.56]		¶ And ȝe þat seke seynte Iames · and seintes of Rome	
Bx.5.60 [KD.5.57]		Seketh seynt treuthe · for he may saue ȝow alle	
Bx.5.61 [KD.5.58]		Qui cum patre & filio · þat feire hem bifalle	
Bx.5.62 [KD.5.59]		Þat suweth my sermon · and þus seyde resoun	
Bx.5.63 [KD.5.60]		¶ Thanne ran repentance · and reherced his teme	
Bx.5.64 [KD.5.61]		And gert wille to wepe · water with his eyen	
Bx.5.65 [KD.5.62]		¶ Peronelle proude-herte · platte hir to þe erthe	
Bx.5.66 [KD.5.63]		And lay longe ar she loked · and lorde mercy cryed	
Bx.5.67 [KD.5.64]	Superbia ·^{T}	And byhiȝte to^{T} hym · þat vs alle made	
Bx.5.68 [KD.5.65]		She shulde vnsowen hir serke · and sette þere an heyre	
Bx.5.69 [KD.5.66]		To affaiten hire flessh · þat fierce was to synne	
Bx.5.70 [KD.5.67]		Shal neuere heiȝe herte me hente · but holde me lowe	
Bx.5.71 [KD.5.68]		And suffre to be myssayde · and so did I neuere	
Bx.5.72 [KD.5.69]		But now wil I meke me · and mercy biseche	
Bx.5.73 [KD.5.70]		For al^{T} þis I haue · hated in myne herte	
Bx.5.74 [KD.5.71]	Luxuria ·	**¶ Þanne lecchoure seyde allas · and on owre lady he cryed**	
L.5.74 [KD.5.71]		¶ Þanne lecchoure seyde allas · and on owre lady he cryed	
M.5.74 [KD.5.71]	Luxuria	Þanne lecchour{~~e~~} saide allas . and on owre lady he criede	
Cr1.5.73 [KD.5.71]	Lechoure	Than Lechoure sayd alas , & on our lady he cryed ,	
W.5.74 [KD.5.71]		C ¶ Thanne lechour seide allas . and on oure lady he cryde	
Hm.5.74 [KD.5.71]		¶ Thanne seyde lechour{~~e~~} alas ⁖ and on oure lady cryede ·	Lu{~~r(?)~~}xuria
C.5.74 [KD.5.71]		¶ Thanne lecchour saide allas · an on oure lady he cried	Luxuria
G.6.74 [KD.5.71]	luxuria	then lycho{~~u~~}vre seyd alas & on our ladye cryed^{T}	
O.5.74 [KD.5.71]		¶ Þanne lecchour seyde allas ⁖ & on oure lady criede	
R.5.74 [KD.5.71]		¶ Þanne lecchour seyde allas ⁖ and on oure lady he cried	
F.5.73 [KD.5.71]		/ / [¶] Þe Lecclour / Lecc[h]our seyde ofte allas / & to oure Lady he criȝede.	

Digital apparatus for *Piers Plowman: The B-Version Archetype*. Piers Plowman Electronic Archive. <http://piers.chass.ncsu.edu/texts/Bx>.

Multiple output formats



```
count(//body//p)
```

Find: where

```
      <bibl>
        <name>Augustinus</name>,
        <title>De Trinitate</title>,
        Bk. XIV, c. 4, n. 6 (PL 42:XXX CCSL XXX)</bibl>
    </ref>
    dicit iuxta
    <ref>illud <title ref="#ps">Psalmum</title></ref>
    <cit>
      <quote ana="#ps38_7">verumtamen in imagine
        pertransit homo,</quote>
      <bibl>Psalm 38:7</bibl>
    </cit>
    quia haec imago ita deformis et obscura
    facta est, ut poenae nulla sit.
    Ipsa autem mens humana secundum
    substantiam suam integre manet,
    igitur secundum substantiam suam non est
    imago Trinitatis.</p>

  <p xml:id="b1d3qun-immnsm">Item, mens magis esset imago Dei quam h
    unumquodque propter quod et illud magis.
    Consequens falsum, quia homo magis est similis
    Deo quam mens, et exprimitur ab eo, igitur magis est eius imago.
    Haec consequentia patet per distinctionem imaginis quam ponit
    <ref>
      <seg><name ref="#Augustine">Augustinus</name> <title ref="#deD
      <bibl><name>Augustinus</name>, <title>De diversis quaestionibu
        q. 74 (CCSL ?; PL 40:86), "ut in speculo est imago hominis,
    </ref>
    ubi dicit quod: imago est expressa similitudo illius
    cuius est imago.
    Assumptum probatur, quia homo est ens nobilior quam mens
```

Rationes principales

2 Quod non videtur, quia secundum Augustinus in *Sermone communi de uno martyre* "si servasset in se homo bonum quod in illo creavit Deus, id est imaginem suam, semper laudaret dictum non S79va solum lingua sed et vita" etc. Igitur ex opposito: non sic laudat, igitur non servavit in se bonum illud secundum imaginem Dei, et tamen servat mentem suam. Igitur mens humana non est imago Dei.

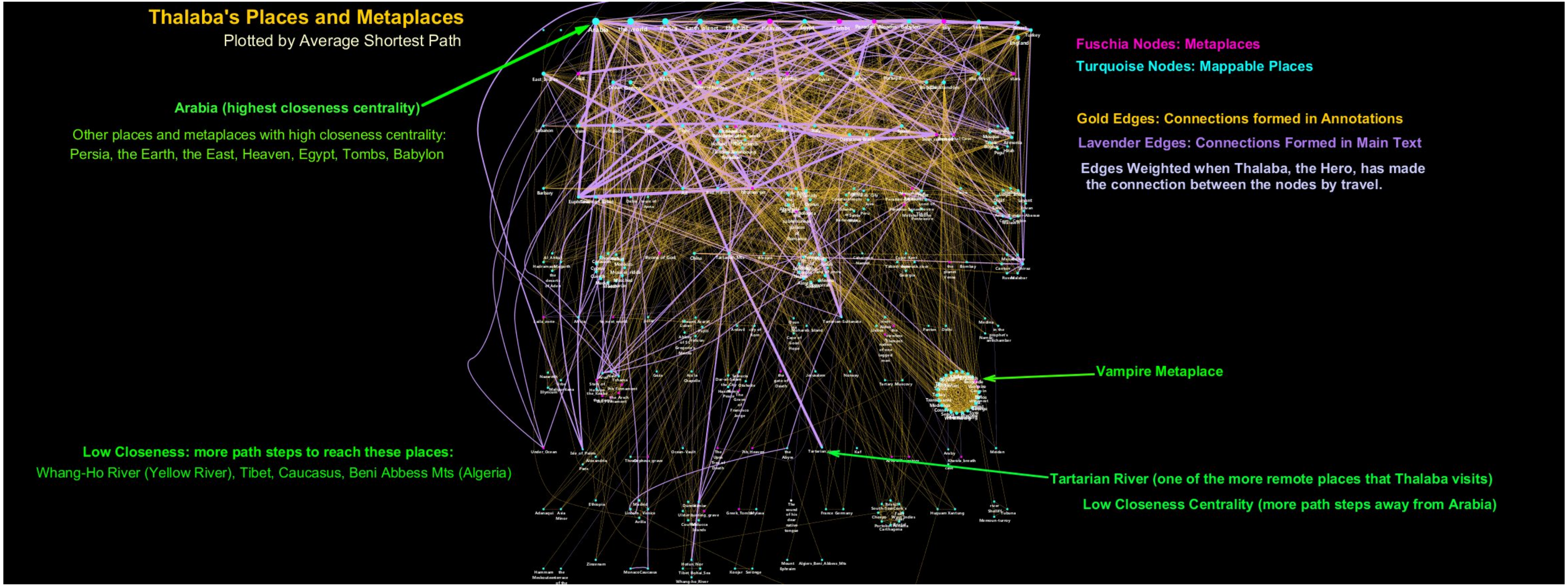
3 Item, 14 *De Trinitate*, c. 7, de parvis dicit iuxta illud *Psalmum* "verumtamen in imagine pertransit homo," quia haec imago ita deformis et obscura facta est, ut poenae nulla sit. Ipsa autem mens humana secundum substantiam suam integre manet, igitur secundum substantiam suam non est imago Trinitatis.

Item, mens magis esset imago Dei quam homo, quia unumquodque propter quod et illud magis. Consequens falsum, quia homo magis est similis Deo quam mens, et exprimitur ab eo, igitur magis est eius imago. Haec consequentia patet per distinctionem imaginis quam ponit Augustinus *De 83 Questionibus*, questione 78, ubi dicit quod: imago est expressa similitudo illius cuius est imago. Assumptum probatur, quia homo est ens nobilior quam mens hominis, igitur est magis similis Deo. Consequentia patet per Anselmum *Monlogion* capitulo 81 et 66, ubi dicit Anselmus "quidquid inter creata Deo constat esse similius, illud necesse est esse natura praestantius." Assumptum probatur quia 1 *Ethicorum* c. 8 in fine, dicit Philosophus quod "felicitas connumerata cum minimo bonorum est eligibilior quam ipsa felicitas, et addit, quod semper maius bonum est eligibilius." Ergo,

5–7 si . . . vita] Auctor incertus (forte Augustinus Hipponensis) *Sermo* CCCXV (PL 39:2349)
10 14 *De Trinitate*, c. 7, de parvis] Augustinus, *De Trinitate*, Bk. XIV, c. 4, n. 6 (PL 42:XXX CCSL XXX)
11 "verumtamen in imagine pertransit homo,"] Psalm 38:7

Demo video for Lbp-Print tool created by Jeffrey Witt for LombardPress <http://lombardpress.org/>.

Sophisticated analysis and display (maps, graphs, etc.)



Thalaba's Places and Metaplaces

Plotted by Average Shortest Path

Arabia (highest closeness centrality)

Other places and metaplaces with high closeness centrality:
Persia, the Earth, the East, Heaven, Egypt, Tombs, Babylon

Low Closeness: more path steps to reach these places:
Whang-Ho River (Yellow River), Tibet, Caucasus, Beni Abbess Mts (Algeria)

Fuschia Nodes: Metaplaces

Turquoise Nodes: Mappable Places

Gold Edges: Connections formed in Annotations

Lavender Edges: Connections Formed in Main Text

Edges Weighted when Thalaba, the Hero, has made the connection between the nodes by travel.

Vampire Metaplace

Tartarian River (one of the more remote places that Thalaba visits)

Low Closeness Centrality (more path steps away from Arabia)

From Elisa E. Beshero-Bondar, "Juxtapositions of Place in Thalaba the Destroyer" <http://ebeshero.github.io/thalaba/index.html>.

XML and Markup

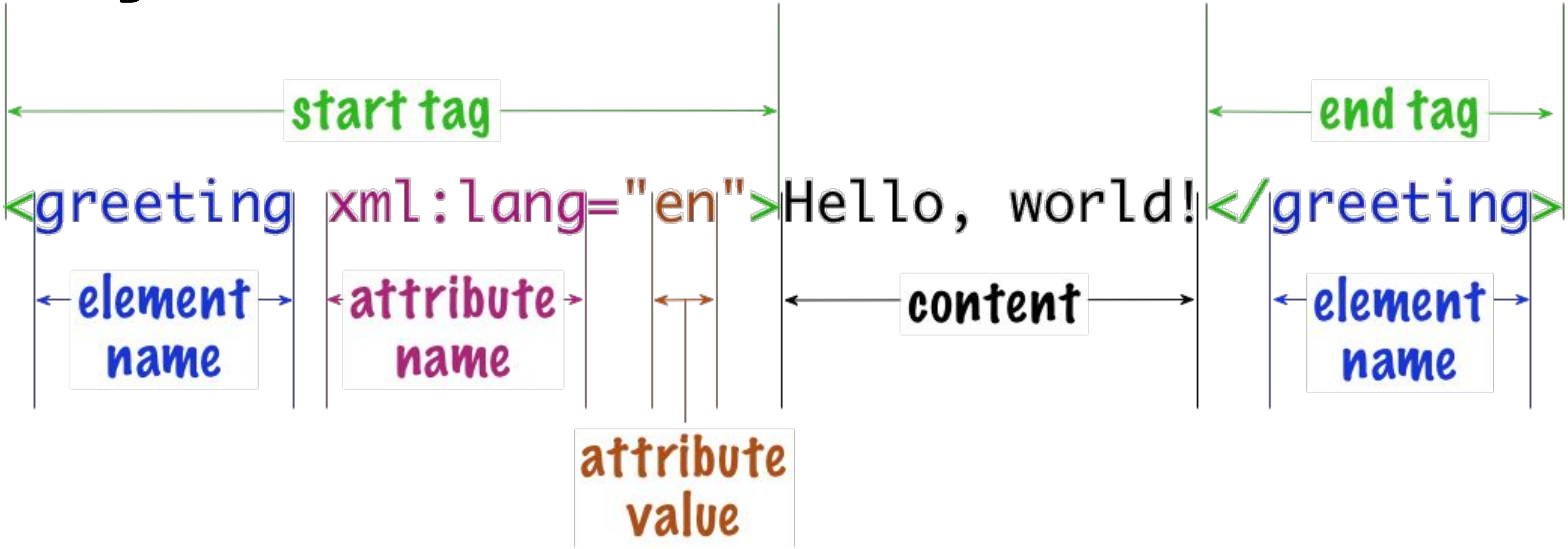
What is markup?

- TEI: "we define *markup*, or (synonymously) *encoding*, as any means of making explicit an interpretation of a text" (*Guidelines* v.1).
- Flavors of markup: proofreaders' marks; paragraphing and punctuation marks; critical apparatus; your underlining and notes in the margin.
- McGann: "there is no such thing as an unmarked text, and the markup systems laid upon documents to facilitate computerized analyses are marking orders laid upon already marked up material" (*Radiant Textuality* 138).
- Computer markup: formal codes for embedding information in a document (for example about structure or formatting); facilitates interchange, processing of unfamiliar documents.

What is XML?

- XML: eXtensible Markup Language
- Privileges the meaning/structure of information over presentation (how it looks)
- Metalanguage: describes other languages that can be defined for specific purposes (hence *extensible*: not limited to preset tags)
- Separates content from metadata (at least in principle)
- Provides a mechanism for *validation*: making sure the document follows the rules
- Allows for the exchange of a wide variety of data between different systems

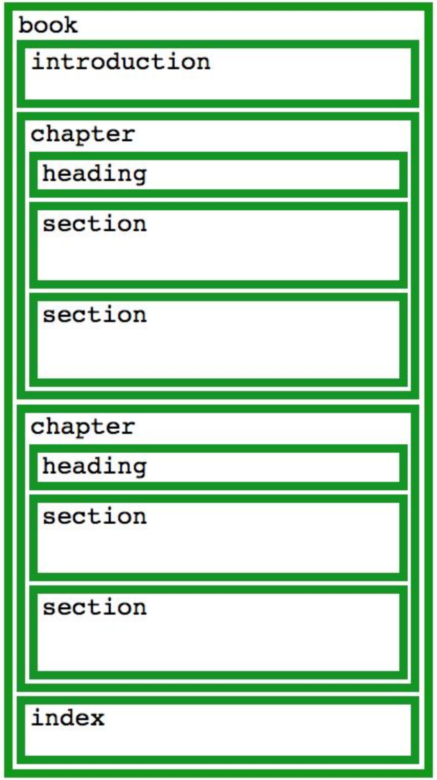
Tags, Elements, and Attributes

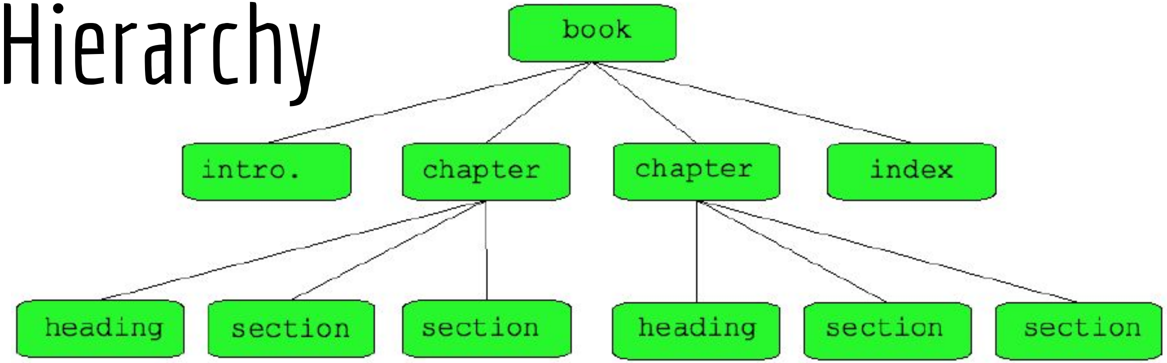


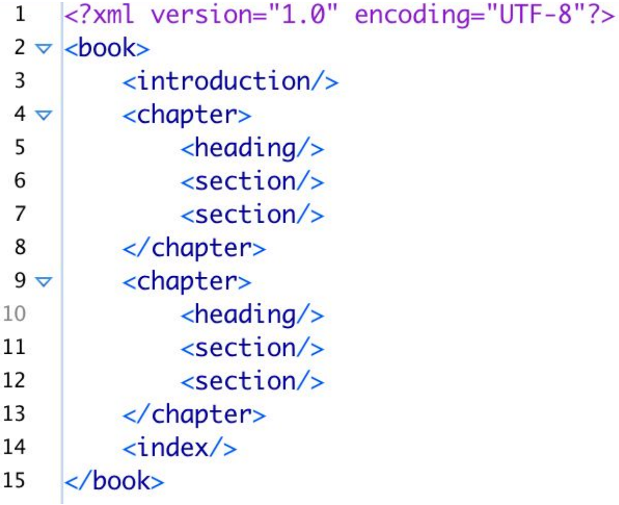
An XML element with no content is called an **empty element** and can be expressed with a single **self-closing tag**: `<tag attribute="value"/>`

Source: Michelle Dalmau and John Walsh, "Text Encoding Initiative Workshop: Intro to Text Encoding," Indiana University <http://dcl.slis.indiana.edu/teiworkshop/teiworkshop_i.pdf>.

XML as Ordered Hierarchy







```xml
<?xml version="1.0" encoding="UTF-8"?>
<book>
    <introduction/>
    <chapter>
        <heading/>
        <section/>
        <section/>
    </chapter>
    <chapter>
        <heading/>
        <section/>
        <section/>
    </chapter>
    <index/>
</book>
```

Elements **nest** inside each other—that is, one element can contain another element. All the elements in an XML document are contained within the **root element**. The overall structure of an XML document is that of a **tree**.

Source: Michelle Dalmau and John Walsh, "Text Encoding Initiative Workshop: Intro to Text Encoding," Indiana University <http://dcl.slis.indiana.edu/teiworkshop/teiworkshop_i.pdf>.

The TEI

What is the TEI?

- TEI = Text Encoding Initiative
- TEI is an organization, the TEI Consortium, which develops standards for digital representation of texts. Founded 1987.
- TEI Guidelines define an encoding language, currently expressed in terms of XML. (When I refer to the TEI, I typically mean this encoding format.)

What does TEI encoding do?

- Allows (primarily textual) documents to be encoded with a high degree of descriptive precision.
    - Not just a transcription: appearance of the document, logical structure, information contained within it, etc.
- Provides a set of elements and attributes that can be used to describe the components of documents, as well as rules for *validating* documents.
- Shared standards enable exchange, interoperability.
- Suite of tools that make it simple to transform and publish TEI documents in multiple formats.
- Flexible and customizable: can define specific, documented variations to meet the need of particular projects.

General Approaches to TEI Encoding

- Organize documents around their logical structure: division into meaningful sections, hierarchies.
    - For example, hierarchy of a book likely to be organized around chapters and paragraphs rather than pages and lines.
    - **But** there are also approaches that prioritize arrangement on the page; flexible to needs of project, material.
- Encode/describe document as it exists, not as you want it to appear.
    - Precise mechanisms for describing formatting of input; formatting of output is left to mechanisms beyond the TEI file.
- Use as much or as little of the TEI tag set as fits the project.
- There are many equally valid ways to encode the same document depending on your interpretation, what features you want to emphasize/analyze.

Hands-On With TEI

TEI Web Editor

- Web app developed by Jeffrey Witt (Assistant Professor of Philosophy, Loyola University Maryland)
- Integrated with GitHub (online saving), Mirador IIIF viewer (show images in editing window)
- Checks if your work is well-formed (does it follow the syntax of XML), but ***not*** if it validates (does it follow the rules of TEI)
- Live preview of your work, lightly formatted
- Online demo (may not be permanent) at https://tei-web-editor.herokuapp.com.

Workshop Instructions

For instructions on completing the workshop exercises, visit https://paulbroyles.github.io/tei-workshop-instructions/.

Where do we go from here?

Learning about TEI

- TEI by Example (http://teibyexample.org/)
- TEI-L Listserv – best for specific questions after you've consulted other resources (https://listserv.brown.edu/archives/cgi-bin/wa?A0=TEI-L)
- The TEI Guidelines (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/)
- Consult with me – pabroyle@ncsu.edu

Reading the TEI Guidelines

- Organized into chapters describing modules that serve a specific purpose (usually, a type of material or approach to encoding).
- Chapters are long and detailed; use the table of contents (and search) to find what you want. You don’t need to read everything!
- Appendix C, Elements, and D, Attributes, are especially useful for getting definitions of particular elements and figuring out where they can go and what they can do.

Software for XML and TEI

- XML is a text-based language, can be edited in any text editor (e.g. Atom, BBEdit [Mac], Notepad [Windows]).
- Dedicated software for editing XML can help avoid errors by checking that XML is well-formed (correct syntax) and valid (follows the rules defined by TEI).
- Commonly used in TEI community: oXygen (http://oxygenxml.com/). Commercial software. Available in CHASS labs.
- oXygen includes built-in support for TEI: provides template for new TEI documents. Also built-in support for transformations with XSLT.

Publishing TEI Editions

- The simple way: CETEIcean
  - What TEI Web Editor uses to publish your work on GitHub Pages
  - JavaScript library to make TEI documents displayable in web browser
  - Create very simple HTML page referencing both the TEI document and CETEIcean JS
  - Can build more sophisticated experiences with CSS to style, additional JavaScript
- The old way: TEI Boilerplate
  - Similar to CETEIcean, but uses a language called XSLT to transform TEI into HTML in the web broswer
  - May not have a long-term future and does not support current versions of XSLT
  - Today, CETEIcean almost always a better choice

Publishing TEI Editions: More Sophisticated Tools

- TEI Publisher
    - Tool for building sophisticated web apps for TEI editions
    - Easy support for multiple documents, search, navigation, etc.
    - Deep formatting control, multiple outputs (PDF, ePub, etc.)
    - Much more complicated setup than CETEIcean, especially for single document; requires more advanced hosting (runs on eXist database).
- EVT (Edition Visualization Technology)
    - Tool for displaying diplomatic editions in parallel with facsimile images
    - Simpler to set up, host than TEI Publisher, but still much more complicated than CETEIcean
    - Main support for diplomatic editions, though a beta version promises support for critical editions.

Analyzing TEI Texts

- Juxta Commons (text collation tool) accepts TEI XML as both an input and an output format; can aid in the editing process or be used to analyze differences among witnesses/versions of a text
- Voyant Tools (text analysis tool) is TEI-aware, will allow you to upload a file in TEI or other XML vocabulary and identify title, text, etc.
- Because TEI files consist of structured data, can be processed to use with any software. Marking up features of interest and providing metadata means you can extract things you can't from plaintext: frequency of correction, how often a person is referenced even if not by name, etc.

Publishing and Transforming TEI: XSLT

- XSLT (eXtensible Stylesheet Language Transformations): a language for transforming XML documents (itself written as a form of XML)
- Can be used to make systematic changes to XML, or to convert to another format (HTML, JSON)
- Many TEI projects use XSLT on the server to create custom interfaces
- XSLT (and the related language XQuery) can be used to process TEI files, gathering data from them directly or turning into another format to use with text processing software
- Sophisticated and complicated programming language, not needed for basic publishing, but worth learning if you want to do deep work